



PEPSI-DOCK

A Detailed Data-Driven
Protein-Protein Interaction Potential
Accelerated By Polar Fourier Correlations

MACARON Workshop- March 21st 2017

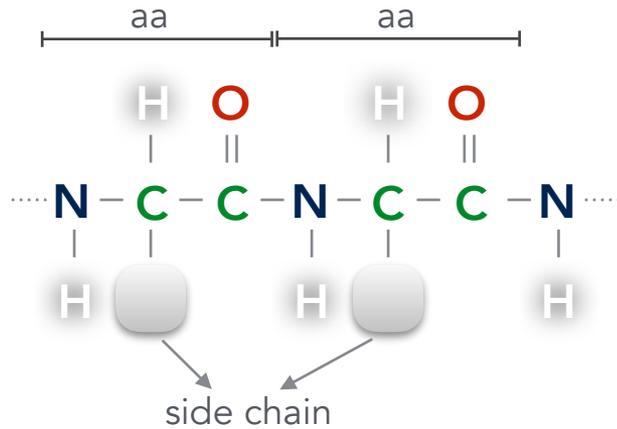
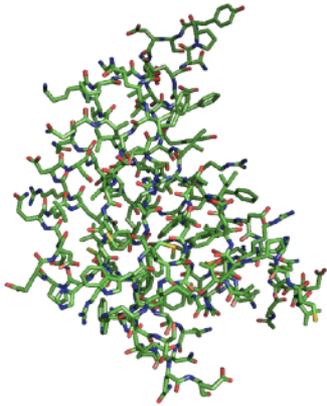
Emilie Neveu, Dave Ritchie, Petr Popov, Sergei Grudin
Nano-D & Capsid INRIA Teams

ABOUT PROTEINS

Definition

protein

(long) chain of amino acids (aa)



20 possible aa

A GUIDE TO THE TWENTY COMMON AMINO ACIDS

AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20 ESSENTIAL AMINO ACIDS (WHICH CAN BE OBTAINED FROM THE DIET). NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESIZED IN THE BODY.

Chart Key: ● ALIPHATIC ● AROMATIC ● ACID ● BASIC ● HYDROPHILIC ● SULFUR CONTAINING ● BASIC ● NON ESSENTIAL ○ ESSENTIAL

Chemical Structure	NAME	CLASS	ESSENTIALITY
<chem>CC(N)C(=O)O</chem>	ALANINE	Aliphatic, Acid	Non-essential
<chem>CC(C)C(N)C(=O)O</chem>	VALINE	Aliphatic, Acid	Essential
<chem>CC(C)C(N)C(=O)O</chem>	ISOLEUCINE	Aliphatic, Acid	Essential
<chem>CC(C)C(N)C(=O)O</chem>	LEUCINE	Aliphatic, Acid	Essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	PHENYLALANINE	Aromatic, Acid	Essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	TRYPTOPHAN	Aromatic, Acid	Essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	TYROSINE	Aromatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	ASPARTIC ACID	Aliphatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	GLUTAMIC ACID	Aliphatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	AMINO ACID	Aliphatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	HISTIDINE	Aliphatic, Acid	Essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	LYSINE	Aliphatic, Acid	Essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	SERINE	Aliphatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	THREONINE	Aliphatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	CISTEINE	Aliphatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	METHIONINE	Aliphatic, Acid	Essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	ALPHA AMINO ACID	Aliphatic, Acid	Non-essential
<chem>C1=CC=C(C=C1)C(N)C(=O)O</chem>	BETA AMINO ACID	Aliphatic, Acid	Non-essential

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between alpha-amino acids and gamma-amino acids is difficult. In these cases, the codes aa(G) and gh(G) are respectively used.

© COMPAGNIE INTERNET 2014 - WWW.COMPAGNIEINTERNET.COM | Twitter: @compagnieinternet | Facebook: www.facebook.com/compagnieinternet | Shared under a Creative Commons Attribution-NonCommercial-NoDerivatives license

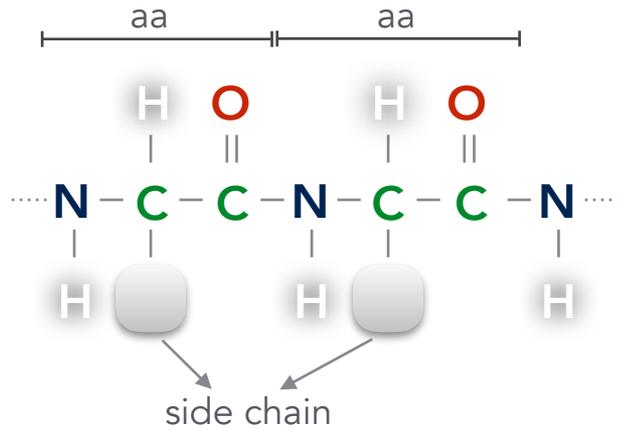
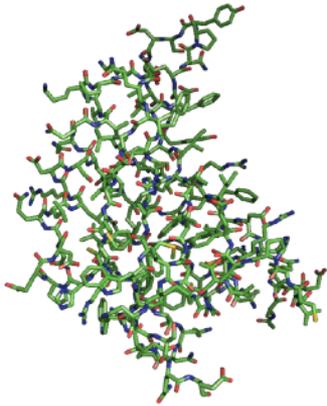
ABOUT PROTEINS

Definition

protein

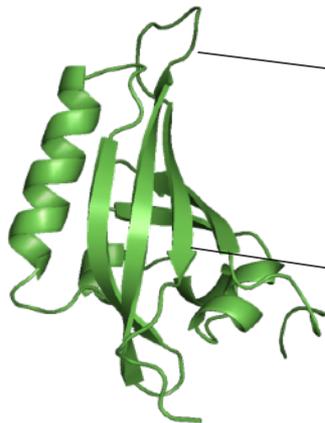
(long) chain of amino acids (aa)

20 possible aa



Representation

"cartoon"-like 3D structure



flexible pieces: structures not well-defined

stable pieces: helices, parallel sheets

ABOUT DOCKING

Structure Prediction

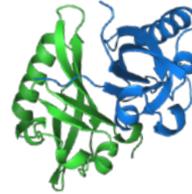
Protein - receptor



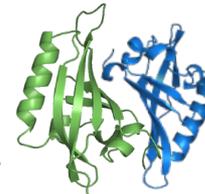
Protein - ligand



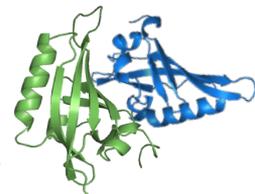
1.



2.

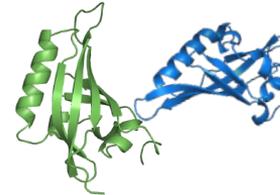


3.



...

N.



ABOUT DOCKING

Structure Prediction

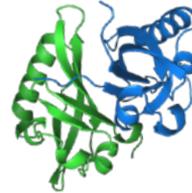
Protein - receptor



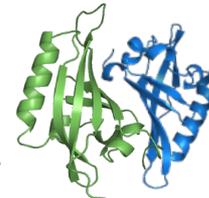
Protein - ligand



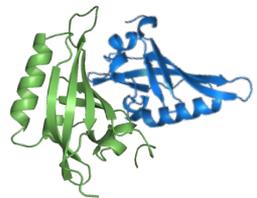
1.



2.

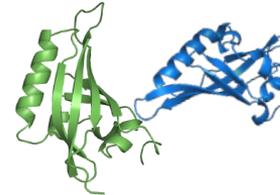


3.



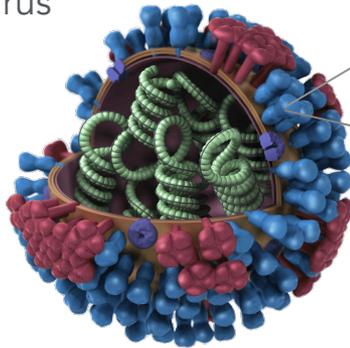
...

N.

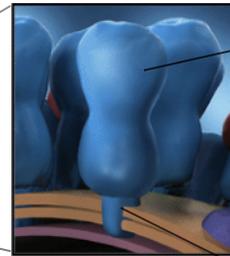


Why so important?

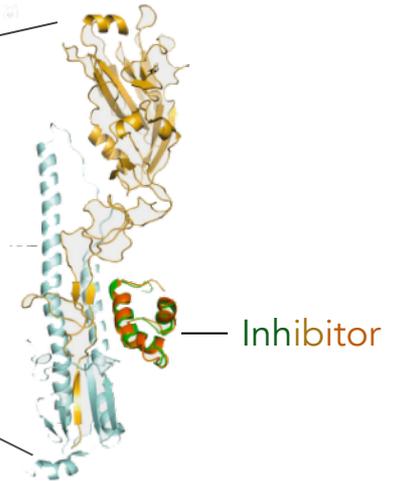
Influenza virus



100 nm



Hemagglutinin protein



Inhibitor

Whitehead, Timothy A., et al. Nature biotechnology 30.6 (2012)

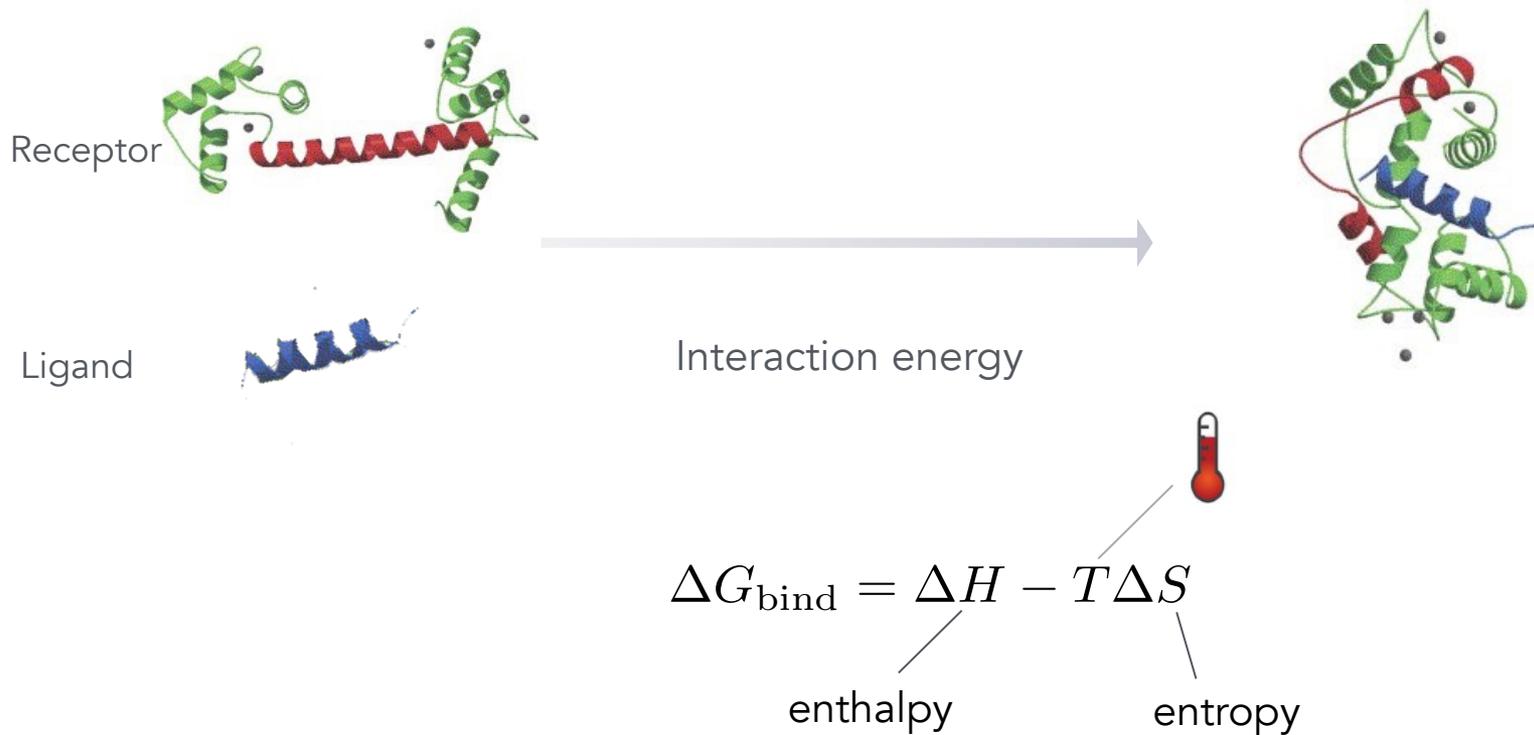
ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

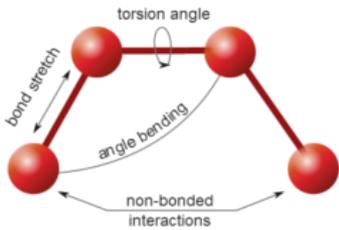
1. Interaction energy to score/assess the structures



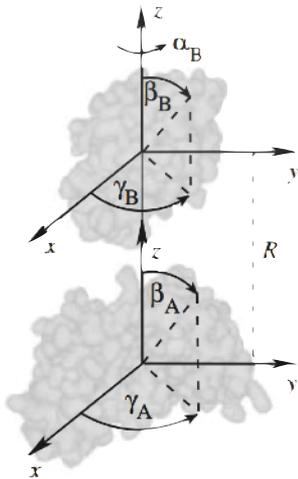
ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

1. **Interaction energy** to score/assess the structures



2. **Search algorithm** + set of parameters



ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

1. **Interaction energy** to score/assess the structures

2. **Search algorithm** + set of parameters



3. **Multilevel approach**: selection of top solutions ; restart with higher resolution

ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

1. **Interaction energy** to score/assess the structures

2. **Search algorithm** + set of parameters

3. **Multilevel approach**: selection of top solutions ; restart with higher resolution

starring

ZDock zdock.umassmed.edu

HexDock hex.loria.fr/hex.php

ClusPro cluspro.bu.edu

AutoDock autodock.scripps.edu

RosettaDock rosie.rosettacommons.org/ligand_docking

DOCK dock.compbio.ucsf.edu

and many others.....

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

GOAL: To improve the first level : large and global search space

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

GOAL: To improve the first level : large and global search space

Simple but accurate interaction energy approximation

- ▶ SVM-based algorithm to learn the atomistic potentials
- ▶ physically interpretable features:
 - number densities of site-site pairs at a given distance
- ▶ arbitrarily shaped atomistic distance dependent interaction potentials

Popov, P., & Grudinin, S. (2015). J. Chem. Info. Model.

Knowledge of Native Protein-Protein Interfaces Is Sufficient To Construct Predictive Models for the Selection of Binding Candidates.

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

GOAL: To improve the first level : large and global search space

Simple but accurate interaction energy approximation

Fast exploration

- ▶ rigid bodies assumption
- ▶ spherical Fourier correlation: complexity from $O(N^9)$ to $O(N^6 \log N)$

D.W. Ritchie, D. Kozakov, and S. Vajda, Hex code

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

GOAL: To improve the first level : large and global search space

Simple but accurate interaction energy approximation

Fast exploration

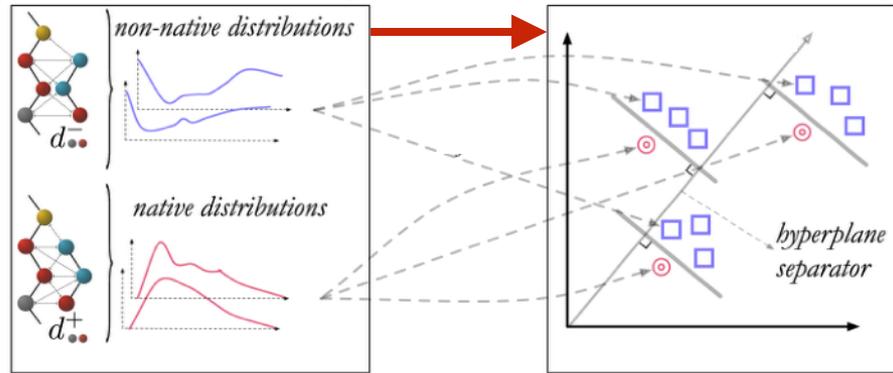
Sparse representation in Gauss-Laguerre basis

1 Features extraction

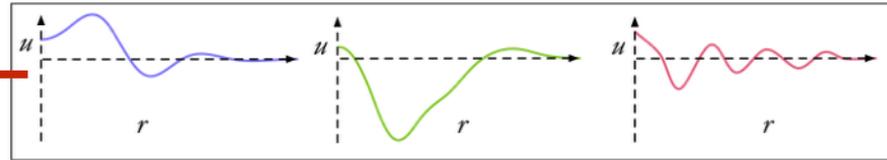
2 Sparse Representation

3

Optimisation



4 Stored 210 atomistic distance dependent potentials

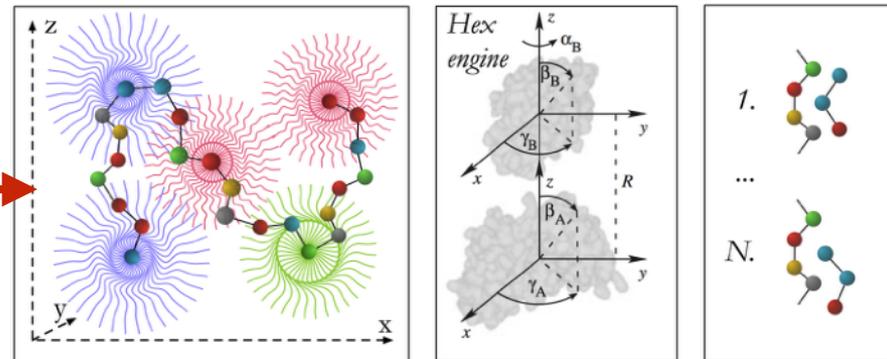


5 From 1D to 3D

6 Fast exploration of the search space

7

Ranked docking predictions

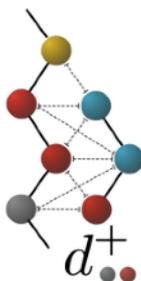


Detailed description of 1-D interactions at the interface

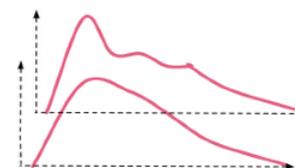
195 native non-redundant complexes

from ITScore Training Set

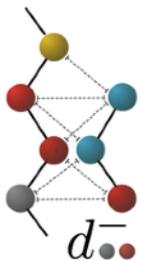
[Zou Lab, University of Missouri Columbia]



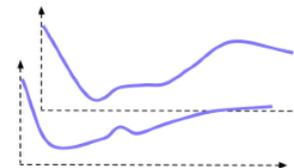
1-D native distributions of atom pairs / distance



40 000 generated false complexes



1-D non-native distributions of atom pairs / distance

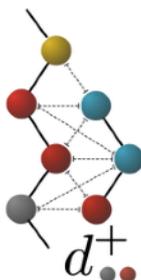


Detailed description of 1-D interactions at the interface

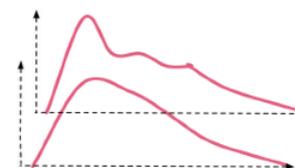
195 native non-redundant complexes

from ITScore Training Set

[Zou Lab, University of Missouri Columbia]

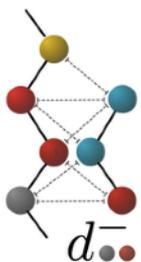


1-D native distributions of atom pairs / distance

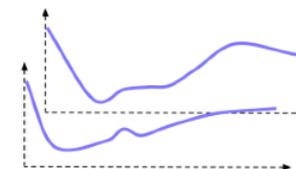


20 different atom types \Rightarrow 210 interactions

40 000 generated false complexes



1-D non-native distributions of atom pairs / distance

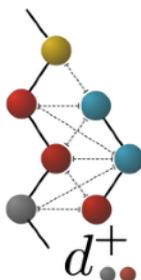


Detailed description of 1-D interactions at the interface

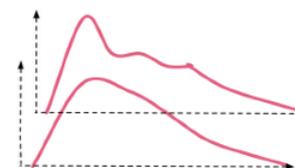
195 native non-redundant complexes

from ITScore Training Set

[Zou Lab, University of Missouri Columbia]

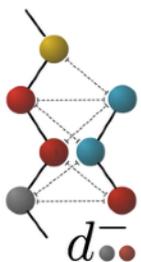


1-D native distributions of atom pairs / distance

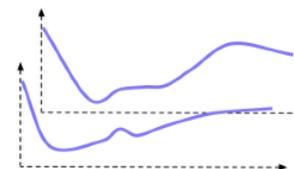


20 different atom types \Rightarrow 210 interactions

40 000 generated false complexes



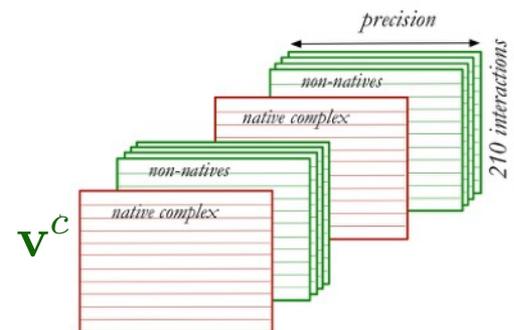
1-D non-native distributions of atom pairs / distance



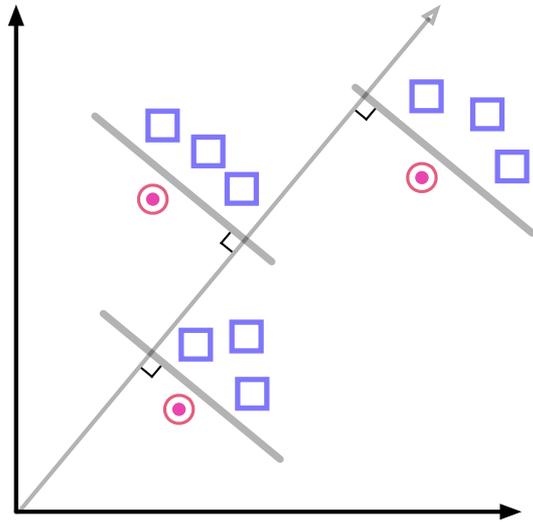
Sparse representation

in a Gauss-Laguerre polynomial basis
scaled to describe distributions up to 30 Å

about 6300 geometric features for each native and non-native complex



Optimal discrimination between native and non native interfaces



features \mathbf{v}^c ; classifier **known**

● native complexes $y^c = 1$

□ associated false complexes $y^c = -1$

hyperplane separator **estimated**

\mathbf{w} normal vector: 1-D interaction potentials

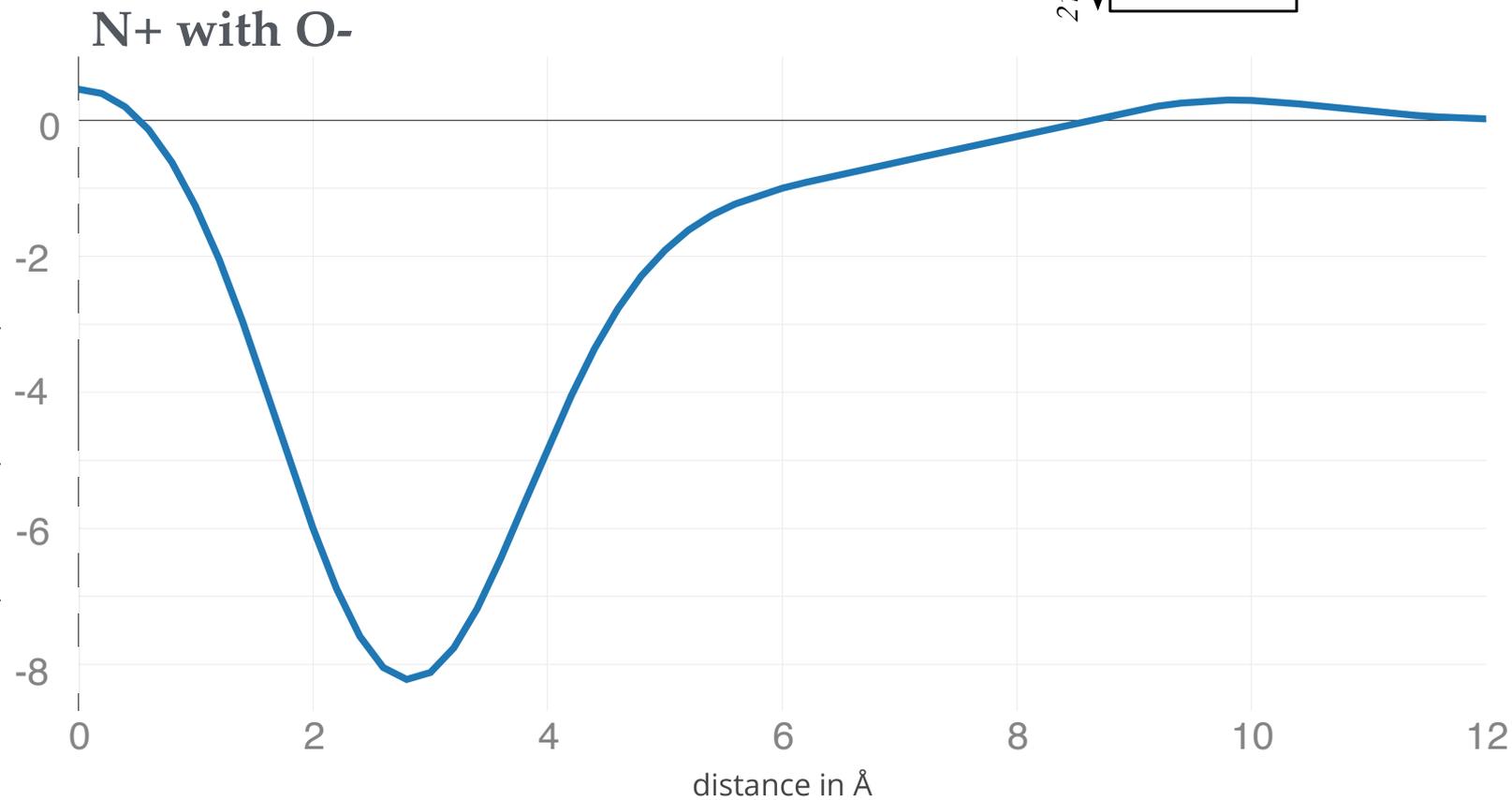
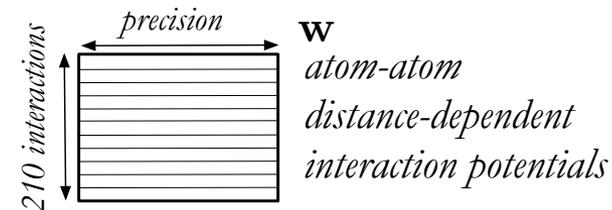
\mathbf{b}^c margin

Convex optimisation problem: Find \mathbf{w} and \mathbf{b}^c that minimise

$$\min_{\mathbf{w}, \mathbf{b}^c} \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{\text{prevents overfitting}} + \underbrace{\gamma \sum_c \log \left(1 + e^{y^c (\mathbf{w}^T \mathbf{v}^c + \mathbf{b}^c)} / \gamma \right)}_{\text{penalises misclassification}}$$

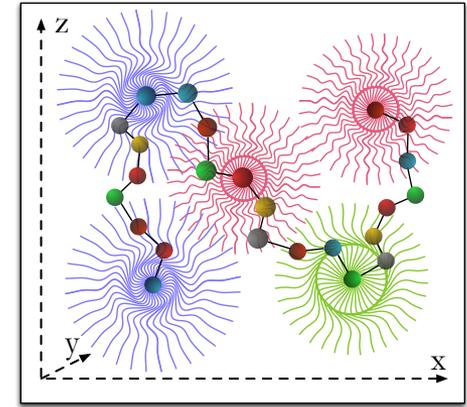
Knowledge of Native Protein–Protein Interfaces Is Sufficient
To Construct Predictive Models for the Selection of Binding Candidates.

Popov, Grudinin, 2015, *J Chem Info Model*.



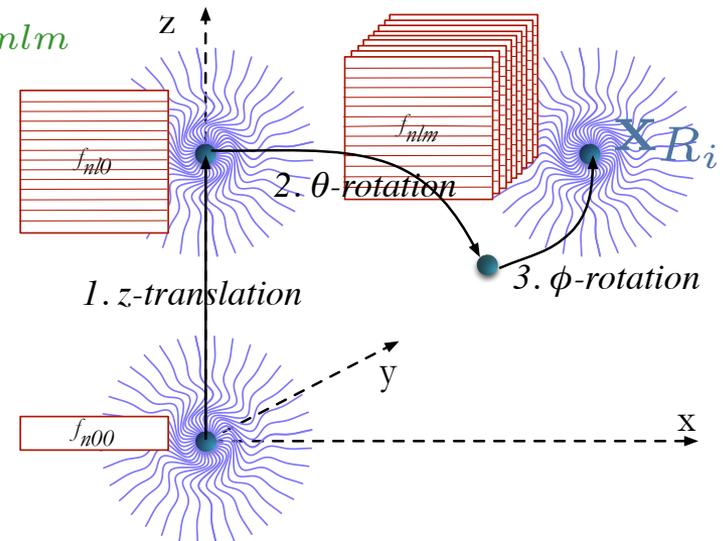
Linear sum of atom-atom convolution with potentials and densities

$$E = \sum_{\text{pairwise interactions } ij} \sum_{R_i} \sum_{L_j} \iiint_V f_{ij}(\mathbf{x} - \mathbf{x}_{R_i}) g(\mathbf{x} - \mathbf{x}_{L_j}) dV$$



Representation with truncated polynomial expansion

$$\iiint_V f_{ij}(r) g(r - \mathbf{x}_{L_j}) dV = \sum_{nlm} \underbrace{(\mathbf{R} \cdot \mathbf{T} \cdot \mathbf{W})_{nlm}}_{= \mathbf{f}_{nlm}^{ij}} \cdot \mathbf{g}_{nlm}$$



Rigid body assumption

Energy depends to rigid positions of proteins

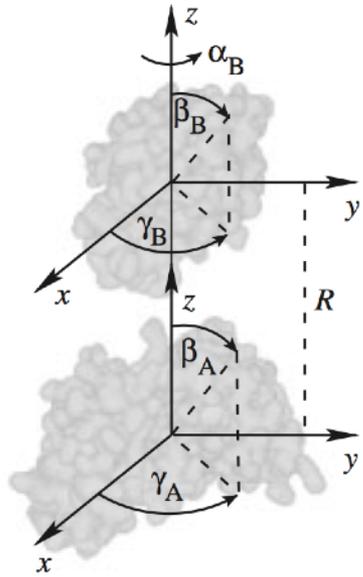
$$E(R, \beta_A, \gamma_A, \beta_B, \gamma_B, \alpha_B)$$

- ▶ 1 translation and 5 rotations to adjust
- ▶ discretised to enable exhaustive search

$$R \in [0 : 1 : 40 \text{ \AA}]$$

$$\alpha \in [0 : 7.5 : 360^\circ]$$

$$(\beta, \gamma) \in [0 : 7.5 : 180^\circ]^2$$



Rigid body assumption

Energy depends to rigid positions of proteins

$$E(R, \beta_A, \gamma_A, \beta_B, \gamma_B, \alpha_B)$$

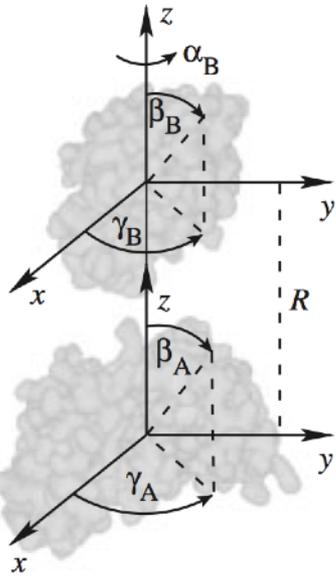
- ▶ 1 translation and 5 rotations to adjust

$$R \in [0 : 1 : 40 \text{ \AA}]$$

$$\alpha \in [0 : 7.5 : 360^\circ]$$

- ▶ discretised to enable exhaustive search

$$(\beta, \gamma) \in [0 : 7.5 : 180^\circ]^2$$



Fast exhaustive search

Truncated expressions using spherical Fourier correlation

$$E(R, \beta_A, \gamma_A, \beta_B, \gamma_B, \alpha_B) = \text{DFT}^{-1} [\text{DFT}(\mathbf{R}_{\beta_A, \gamma_A} \mathbf{T}_R \mathbf{f}_A) \cdot \text{DFT}(\mathbf{R}_{\alpha_B, \beta_B, \gamma_B} \mathbf{g}_B)]$$

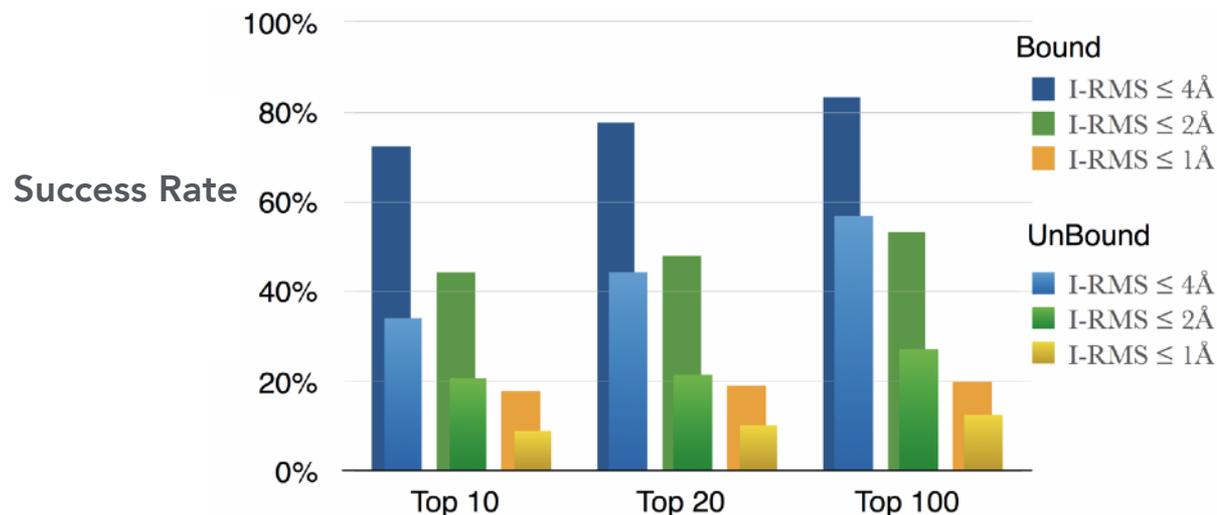
complexity from $O(N^9)$ to $O(N^6 \log N)$: **10^9 poses in ~ 10 min**

Accelerating and Focusing Protein-Protein Docking Correlations Using Multi-Dimensional Rotational FFT Generating Functions.
D.W. Ritchie, D. Kozakov, and S. Vajda (2008). *Bioinformatics*. 24 1865-1873.

Test on 88 complexes from the Docking Benchmark Set v5.0 for which the separation distance ≤ 30 Å

Docking Benchmark Set = the only existing benchmark to compare different docking algorithms

[Hwang, Vreven, Janin, Weng, 2010]

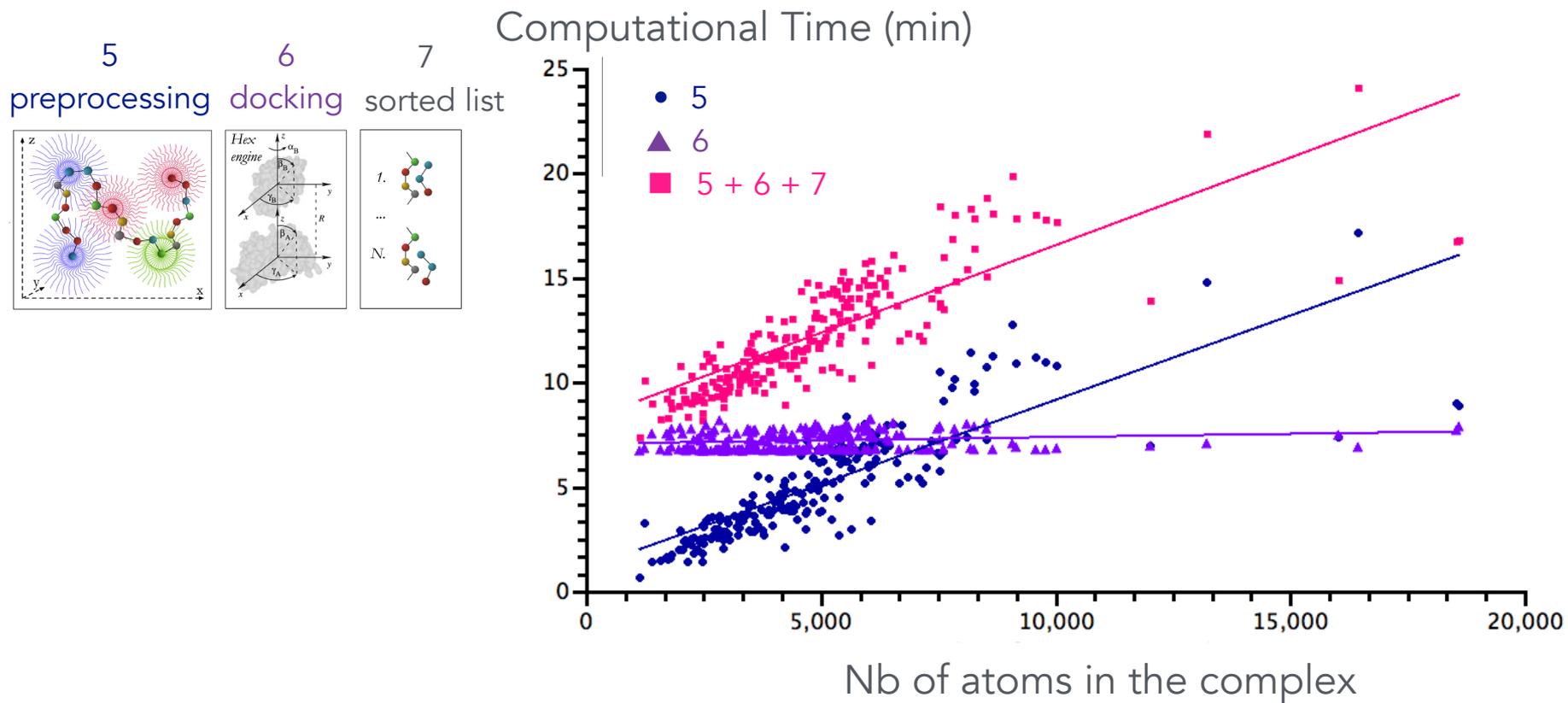


Comparison on v4.0

Top 10 for I-RMS ≤ 2.5 Å

Category (Nb. of Complexes)	ZDOCK.	SwarmDock	PEPSI-Dock
Easy (45)	12	11	13
Medium (15)	1	1	1
Difficult (15)	0	0	0

Running Time of PEPSI-Dock measured on a modern laptop



Docking of 10^9 poses in less than 10 min on a laptop ~ weeks of a 1 μ s MD simulation

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

A docking automatic algorithm for the first stage of the docking pipeline

- » novelty: arbitrarily -shaped + distance-dependent potentials combined with a FFT search sampling technic

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

A docking automatic algorithm for the first stage of the docking pipeline

- » novelty: arbitrarily -shaped + distance-dependent potentials combined with a FFT search sampling technic

TO DO

1. Improve unbound predictions: use other training set

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

A docking automatic algorithm for the first stage of the docking pipeline

- » novelty: arbitrarily -shaped + distance-dependent potentials combined with a FFT search sampling technic
- ▶ Bound sets: High-rank predictions 
- ▶ Large distances  loss of precision
- ▶ Unbound sets: similar results than SwarmDock or ZDOCK
- ▶ Adaptation to other types of interactions   

TO DO

1. Improve unbound predictions: use other training set

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

A docking automatic algorithm for the first stage of the docking pipeline

- » novelty: arbitrarily -shaped + distance-dependent potentials combined with a FFT search sampling technic
- ▶ Bound sets: High-rank predictions 
- ▶ Large distances  loss of precision
- ▶ Unbound sets: similar results than SwarmDock or ZDOCK
- ▶ Adaptation to other types of interactions 

TO DO

1. Improve unbound predictions: use other training set
2. Deal with the docking of large proteins: use other sampling

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

PEPSI-Dock, Neveu *et al.*, Bioinformatics, 2016

PEPSI-DOCK

Polynomial Expansions of Protein Structures and Interactions for Docking

PEPSI-Dock, Neveu et al., Bioinformatics, 2016



<https://www.samson-connect.net>





THANKS

@ Sergei Grudin
@ Petr Popov
Nano-D team
INRIA Grenoble

@ David Ritchie
Capsid Team
INRIA Nancy
ANR

<https://www.samson-connect.net>



ANY QUESTION?